A NO-REFERENCE HDR-VDP 2 METRIC TOWARDS A DEEP-LEARNING APPROACH

Francesco Banterle[1] *Alessandro Artusi*[2] *Alejandro Moreo*[1] *Fabio Carrara*[1]

[1] ISTI-CNR, Pisa, Italy
[2] RISE Ltd, Nicosia, Cyprus

ABSTRACT

HDR-VDP 2 has convincingly shown to be a reliable metric for image quality assessment, and it is currently playing a remarkable role in the evaluation of complex image processing algorithms. However, HDR-VDP 2 is known to be computationally expensive (both in terms of time and memory) and is constrained to the availability of a ground-truth image (the so-called *reference*) against to which the quality of a processed imaged is quantified. These aspects impose severe limitations on the applicability of HDR-VDP 2 to real-world scenarios involving large quantities of data or requiring real-time responses. To address these issues, we propose *Deep No-Reference Quality Metric* (DNR-QM), a deep-learning approach that learns to predict the global image quality feature (i.e., the *mean-opinion-score* index *Q*) that HDR-VDP 2 computes. DNR-QM is no-reference (i.e., it operates without a ground truth reference) and its computational cost is substantially lower when compared to HDR-VDP 2 (by more than an order of magnitude). We demonstrate the performance of DNR-QM in a variety of scenarios, including the optimization of parameters of a denoiser and JPEG-XT.

Index Terms— HDR-VDP 2, Objective metric, No-reference, Deep-learning, Image quality assessment.

1. INTRODUCTION

In computer vision and related tasks, the quality of synthetic images is commonly assessed either through user studies or through objective metrics. Despite the former being more reliable, the large number of users and images typically required by subjective studies makes the adoption of objective metrics more attractive in applicative scenarios. For this reason, the community is devoting a great deal of effort to the study of new sophisticated objective metrics [1]. Objective metrics are nonetheless very reliable, and especially so when focusing on the simulation of complex aspects of the *Human Visual System* (HVS); something which, however, comes at a high computational cost. Yet another limitation is to be found in the so-called Fully-Reference (FR) metrics, a class of objective metrics that, as their name suggests, require the availability of a ground truth image (i.e., a version of the image being evaluated that contains no artifacts) — an unaffordable price for many practical scenarios. For all these reasons, the study of reliable No-Reference (NR) metrics is nowadays gaining considerable research attention.

Despite being very popular as an objective metric in the field, the *High Dynamic Range Visual Differences Predictor* (HDR-VDP 2) [2] is a good example of the aforementioned limitations. Its high computation cost, along with the need for a ground truth reference, precludes HDR-VDP 2 (and related metrics) from being used in several quality assessment scenarios such as standardization, real-time quality assessment, etc. This altogether motivates the need for more efficient, yet effective, objective metrics that can predict visual significant differences of test images without relying on ground truth references.

In this paper, we investigate practical solutions to counter these issues. The main contribution of this paper concerns the study and evaluation of a deep-learning-based alternative to the popular HDR-VDP 2 [2] implementation. We propose *Deep No-Reference Quality Metric* (*DNR-QM*), a model which (i) is able to predict visual metric features that are believed to be well correlated with the *mean-opinion-score* (MOS) (e.g., the quality index *Q* of HDR-VDP 2 – [2]), that (ii) does so at a fraction of the time HDR-VDP 2 demands (more than an order of magnitude faster), and (iii) without the need of a ground truth reference.

We tested DNR-QM on a variety of scenarios designed to demonstrate its robustness and flexibility. DNR-QM performs in real-time and is thus suitable to be integrated as the main optimization component into different applications, which we tested in our experiments. These include a denoiser for Standard Dynamic Range (SDR) images, and a *High Dynamic Range* (HDR) encoder. Finally, our framework has very low computational costs. This would be very helpful for standardization bodies (e.g., JPEG and MPEG) to use it for extremely large datasets.

2. RELATED WORK

When the original image is available, it can directly be used by Fully-Reference (FR) metrics to assess the quality of a processed (e.g., distorted, compressed) version of it. In such cases, the original image is typically called the *ground truth* or the *reference*, meaning that the quality score (whichever

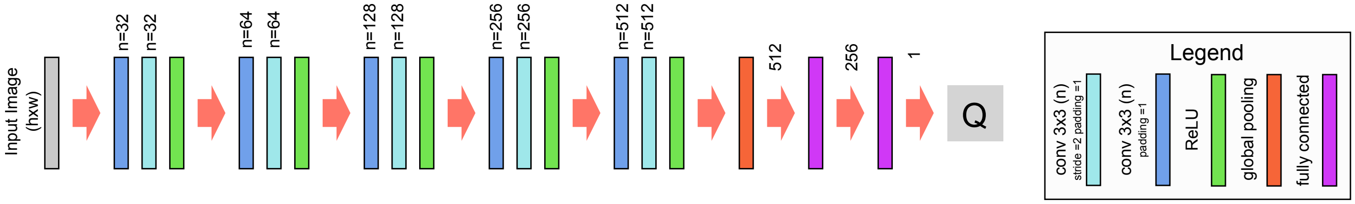


Fig. 1. The proposed architecture for computing HDR-VDP 2.

the definition) should be higher as the processed image approaches the ground truth. However, in many practical applications, it is often the case that a ground truth does not exist or is not accessible. No-Reference metrics (NR), a.k.a. Objective-Blind metrics, are used in such contexts, i.e., metrics that perform Image Quality Assessment (IQA) without any prior knowledge about the original (distortion-free) image.

Several NR metrics have been described in the literature, paying special attention to the difficulties that the absence of a proper reference imposes to the model of evaluation. A possible idea to overcome this limitation is to extract some statistics from the distorted image and compare them against similar statistics previously extracted from natural undistorted images [3, 4, 5]. Another approach consists of extracting specific characteristics of the distortion when the type of artifact is known beforehand. Some approaches following this intuition use local gradient [6], saliency map and Support Vector Regression (SVR) [7], or measure the power of the blocking signal [8].

Convolutional Neural Networks (CNN) are continuously showing impressive performance in many computer vision tasks, and IQA is by no means an exception. Most of the recently proposed CNN-based models focus on the FR case. Among those, [9] compares feature maps extracted from the CNN layers of test and reference images, [10] learns the HVS behavior from the underlying data distribution of IQA databases, and [11] fuses scores obtained from multiple quality indices into one score. Other proposed metrics are data-driven [12], or train a model to learn perceptual transforms [13, 14]. CNN-based approaches that tackle the NR problem also exist. Some examples include purely data-driven approaches [15], approaches that learn rules from linguistic descriptions [16], others that extract different gradient-based features [17, 18], or approaches modeling the perceptual masking effects in distorted HDR images [19].

3. DEEP NO-REFERENCE QUALITY METRIC

Our goal is to predict the quality value (Q) of HDR-VDP 2 [2] given exclusively a distorted image as input (i.e., in the absence of reference). Although HDR-VDP 2 generates additional outputs, such as a *threshold normalized contrast map* (C_{map}) and its maximum value, we restrict our attention to the quality value in this research. The reason why we avoid the prediction of the C_{map} responds to the fact that such maps turn out to be of limited help in applicative scenarios like standardization committees (e.g., JPEG and MPEG) since, when analyzing large datasets [14], a single value is typically preferred.

As the model architecture, we adopted a CNN model [20], since such architectures have provided high-quality results in several computer vision/imaging tasks while at the same time run very efficiently on GPUs. Specifically, we take an already existing architecture [14] as a starting point, which is based on a modified version of the U-Net [21].

We tweaked this network by taking a single image as input (thus removing the part of the network dealing with the reference) and removing the max-pooling operator from the network. The reason behind getting rid of the max-pooling layer is to retain as much information as possible when detecting distortions (pooling strategies are known to discard information). During preliminary experiments, we indeed observed that removing the max-pooling layer brings about a 5% of decrease in the validation loss.

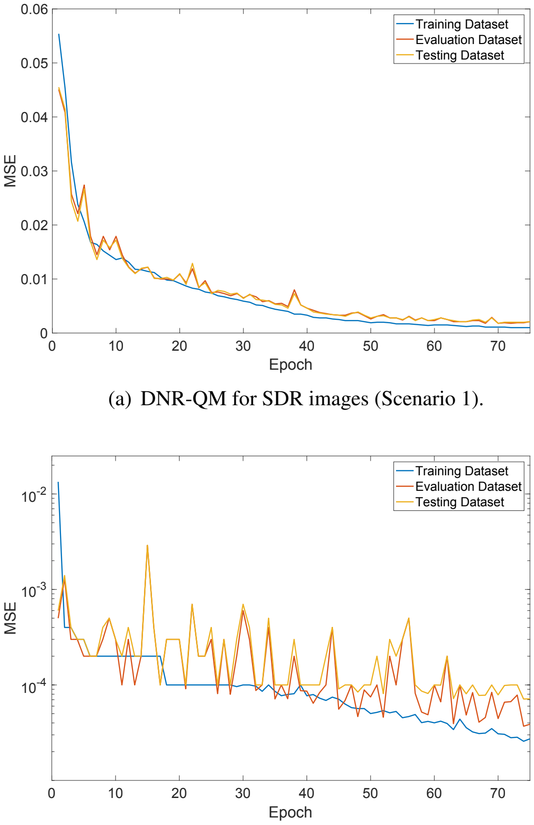
Furthermore, we decided to force the network to work only on luminance images (i.e., a single channel image) instead of having three color channels. This is because we noticed there was not a drop in the quality of the performance. Furthermore, we gained a significant speed-up when evaluating the network, see Section 5.2. Figure 1 shows the scheme of our network.

3.1. Datasets

We employed two datasets by Artusi et al. [14] to train our model (those datasets are available upon request). Since our goal is to have a no-reference HDR-VDP 2, we focused on Scenario 1 and Scenario 2 (as defined in [14]).

Scenario 1 is a dataset of SDR images presenting different distortions (e.g., blur, quantization, noise, etc.) and consisting of 16,002 images. We randomly split the dataset in 80% for training, 10% for validation, and 10% for testing.

Scenario 2 is a dataset of HDR images with different levels of JPEG-XT [22, 23] compression using all profiles, and it has 14,418 images. As for Scenario 1, we randomly split the dataset in 80% for training, 10% for validation, and 10% for testing.



(a) DNR-QM for SDR images (Scenario 1).

(b) DNR-QM for HDR Compression (Scenario 2)

Fig. 2. Plots of the training, evaluation, and testing datasets for each epoch. The minimum Evaluation MSE is 3.6×10^{-5} (Scenario 1) and 1.8×10^{-3} (Scenario 2).

4. TEST CONDITIONS

We trained our model on a Linux machine (Ubuntu 18.04) equipped with an Intel CPU Core i7-7800X (3.50 GHz) with 64 GB of memory and an NVIDIA GeForce GTX 1080 GPU with 8 GB of memory. We implemented DNR-QM using PyTorch 1.3.1 deep-learning framework.

The pixel values from HDR and SDR images are preprocessed differently, following the indications in [14]. SDR images are linearly scaled from the original range $[0, 255]$ to $[0, 1]$ before feeding them to the network. Instead, we apply the logarithm to HDR images:

$$x' = \log_{10}(x + 1) \,, \tag{1}$$

where x is the input pixel value. By doing so, we obtain an equilibrate scale in the positive only real values that is not biased towards large differences in high luminance values [24].

We trained the network using mini-batch stochastic gradient descent and the Adam update rule [25] with the learning rate set to 0.001. We left the rest of the parameters set to their default values; i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$. As in Artusi et al. [14], we defined the loss function to be the *Mean Square Error* (MSE) between the predicted and the true scores. We initialized all network weights following the Xavier initialization [26]. We set the batch size of DNR-QM to 32 samples, which was the largest parameter for which enough memory could be allocated in our NVIDIA GeForce GTX 1080 GPU. The training set is shuffled whenever an epoch is completed to diminish the impact of order-based biases during training. We set the maximum number of epochs to 75; the training time varies from approximately 6 hours to 10 hours depending on the size of the training set. In all cases, the reported results correspond to the models obtaining the minimum loss as measured in the validation set.

5. RESULTS

5.1. Quality of Learning

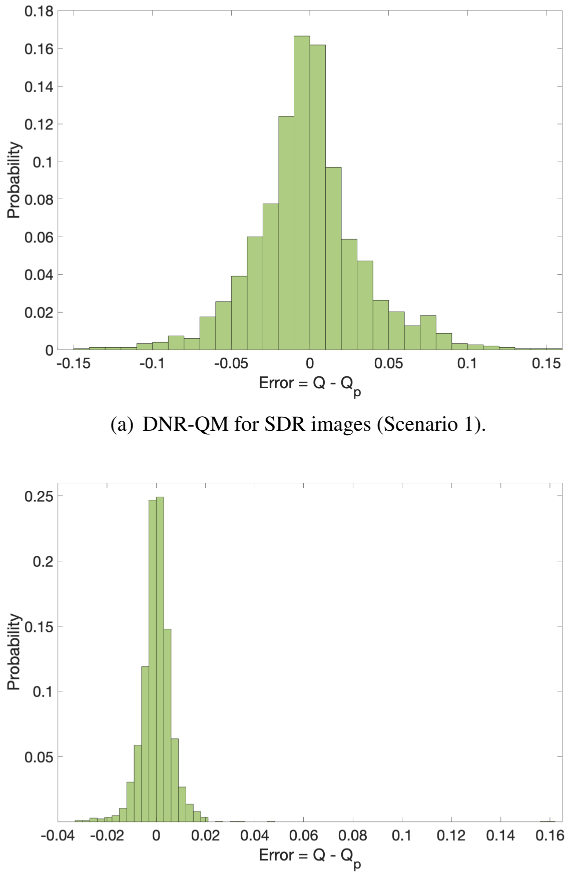
Figure 2 displays the training curves; i.e., the evolution of the loss as evaluated on the training, validation, and testing[1] data. In both scenarios (i.e., Scenario 1 and Scenario 2), the model seems to converge to low loss figures around 75 epochs (no further improvement in validation is observed thereafter).

Figure 3 plots the histograms of errors for the test data with respect to the ground truth for Scenario 1 (a) and Scenario 2 (b). The predictions our model produces are particularly accurate for Scenario 2 (Figure 3 (a)) i.e., for the case of HDR images compressed with JPEG-XT distortions, as witnessed by the narrow distribution of test errors around 0 and the low presence of outliers. Figure 3 (a) shows the model produces comparatively higher errors in Scenario 1 (i.e., for SDR images), but the error remains still withing acceptable margins for many practical applications; in this case, there are no outliers.

5.2. Timings

Figure 4 reports testing times at different image resolutions for our method DNR-QM against the ground-truth HDR-VDP 2 [2] and the deep learning model DIQM [14] (currently, the state of the art for reproducing HDR-VDP 2 using CNNs). The implementation we used for HDR-VDP 2 [2] is the variant proposed in [14] which takes advantage of the GPU parallel processing via CUDA libraries. All times reported are clocked in the same machine (see Section 4 for details). Note the size of the input images ranges from 128×128 (VGA resolution) to 8-Mpixel resolution. The most interesting fact that emerges from Figure 4 is that our method is faster than the competitors. Specifically, it happens to be 1.5 times faster than DIQM [14]. This is because our network works on luminance images (one channel) instead of on RGB (three channels) images as DIQM. Furthermore,

[1] Test data is, of course, assumed unavailable during model training. We still report test trends here for demonstrating purposes.



(a) DNR-QM for SDR images (Scenario 1).

(b) DNR-QM for HDR Compression (Scenario 2).

Fig. 3. The histograms of the error distribution between the ground truth Q value and the predicted value by our network, Q_p, for the testing dataset.

our method is nearly two orders of magnitude faster than (the CUDA based implementation) HDR-VDP 2 which, in turn, is unable to process images larger than 4-MPixel due to high usage of memory.

5.3. Applications

Our proposed metric can be used in different applications in which not only the efficacy of the estimation of the image quality, but also the efficiency, are important. In this section, we demonstrate the performance of DNR-QM in two representative applications.

As a first application, we implemented a denoiser based on the bilateral filter [27]; more sophisticated methods for denoising exist in literature, but it is merely used here as an example. This denoiser uses DNR-QM to optimize the filter parameters; i.e., σ_r (which controls the smoothing threshold) and σ_s (which controls the size of the spatial kernel/neighborhood). Then, we ran it on a 1Mpixel noisy image and we compared the result against DIQM [14]. While both outputs were similar, our method took 1.34s to run, DIQM took 1.82s on the same image.

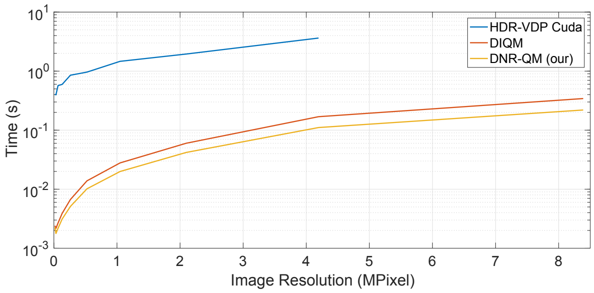


Fig. 4. Testing times for the prediction of the Q value for our method DNR-QM against the ground-truth HDR-VDP 2 [2] (CUDA version) and DIQM [14]. Note the log scale.

The second application is a companding scheme based on a parameterized sigmoid function, $y(x) = ax/(1 + ax)$, for compressing HDR images using JPEG. This application uses the predictions of DNR-QM to optimize the sigmoid parameter a and the JPEG quality parameter for compressing HDR images in an efficient way. We tested it on a 4Mpixel HDR image. Again, both outputs were similar. In this case, our method took 11.67s to run, while DIQM took 18.74s on the same image.

6. CONCLUSIONS AND FUTURE WORK

In this work, we have shown how to convert and distill HDR-VDP 2 into a no-reference metric, that we dub DNR-QM, using a deep learning approach. The results we have obtained show that DNR-QM performs robustly (in terms of error w.r.t. the ground truth) in scenarios concerning both SDR and HDR images. Furthermore, our method is computationally efficient, especially when compared to a CUDA-based implementation of the original HDR-VDP 2.

In future work, we plan to broaden the generalization of the method by incorporating other types of inputs (e.g., including viewing conditions) to our network. This may imply extending the large datasets made available by Artusi et al. [14]. Finally, we want to extend this work by comparing the performance of our network against other non-reference solutions for SDR and HDR distortions.

7. REFERENCES

[1] Alessandro Artusi, Francesco Banterle, Tunç Ozan Aydı n, Daniele Panozzo, and Olga Sorkine-Hournung, *Image Content Retargeting: Maintaining Color, Tone, and Spatial Consistency*, CRC Press, september 2016.

[2] Manish Narwaria, Rafal K. Mantiuk, Mattheiu Perreira Da Silva, and Patrick Le Callet, "HDR-VDP-2.2: A calibrated

method for objective quality prediction of high dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, 2015.

[3] Tomás Brandão and Maria Paula Queluz, "No-reference image quality assessment based on dct domain statistics," *Signal Process.*, vol. 88, no. 4, pp. 822–833, Apr. 2008.

[4] Michele A. Saad, Alan C. Bovik, and Christophe Charrier, "A dct statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, pp. 583–586, 2010.

[5] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[6] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *2009 International Workshop on Quality of Multimedia Experience*, July 2009, pp. 64–69.

[7] Zhongyi Gu, Lin Zhang, and Hongyu Li, "Learning a blind image quality index based on visual saliency guided sampling and gabor filtering," in *ICIP*. 2013, pp. 186–190, IEEE.

[8] Zhou Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, Sep. 2000, vol. 3, pp. 981–984 vol.3.

[9] Amirshahi Seyed Ali, Pedersen Marius, and X. Yu Stella, "Image quality assessment by comparing cnn features between images," *Journal of Imaging Science and Technology*, vol. 60, no. 6, pp. 060410:1–060410:10, 2016.

[10] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] Tsung-Jung Liu, Weisi Lin, and C.-C. Jay Kuo, "Ieee transactions on image processing," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1793 – 1807, 12 2012.

[12] Martin Čadík, Robert Herzog, Rafal Mantiuk, Radoslaw Mantiuk, Karol Myszkowski, and Hans-Peter Seidel, "Learning to predict localized distortions in rendered images," *Computer Graphics Forum*, 2013.

[13] N. Ye, M. Prez-Ortiz, and R. K. Mantiuk, "Trained perceptual transform for quality assessment of high dynamic range images and video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1718–1722.

[14] Alessandro Artusi, Francesco Banterle, Alejandro Moreo, and Fabio Carrara, "Efficient evaluation of image quality via deep-learning approximation of perceptual metrics," *IEEE Transactions on Image Processing*, vol. 29, pp. 1843–1855, oct 2019.

[15] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[16] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 6, pp. 1275–1286, 2015.

[17] Debarati Kundu, Deepti Ghadiyaram, Alan C. Bovik, and Brian L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures," *IEEE Trans. Image Processing*, vol. 26, no. 10, pp. 4725–4740, 2017.

[18] Le Kang, Peng Ye, Yi Li, and David S. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1733–1740.

[19] N. K. Kottayil, G. Valenzise, F. Dufaux, and I. Cheng, "Blind quality estimation by disentangling perceptual and noisy features in high dynamic range images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1512–1525, March 2018.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[22] Alessandro Artusi, Rafal Mantiuk, Thomas Richter, Pavel Korshunov, Philippe Hanhart, Touradj Ebrahimi, and Massimiliano Agostinelli, "JPEG XT: A Compression Standard for HDR and WCG Images [Standards in a Nutshell]," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 118–124, 2016.

[23] Thomas Richter, Walt Husak, , Niman Ajit, Ten Arkady, Pavel Korshunov, Touradj Ebrahimi, Alessandro Artusi, Massmiliano Agostinelli, Shigetaka Ogawa, Peters Schelkens, Takaaki Ishikawa, and Tim Bruylants, "JPEG XT information technology: Scalable compression and coding of continuous-tone still images," .

[24] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 178:1–178:15, Nov. 2017.

[25] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[26] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[27] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Jan 1998, pp. 839–846.